FAKE JOB POSTING PREDICTION

PROJECT REPORT

SUBMITTED TO

VISHWAKARMA UNIVERSITY, PUNE

FOR THE DEGREE

OF

MASTER OF SCIENCE

IN

STATISTICS – BIG DATA ANALYTICS

BY

HAMZA DALAL,

SHRAVANI UNCHE,

BHAGYALAXMI MAHARANA &

DEEKSHA DESHMUKH

Under the supervision of

Dr. Nazia Wahid

Head of Department of Mathematics & Statistics

Faculty of Science and Technology

Vishwakarma University, Pune

Prof. Mahfooz Alam

Assistant Professor

Department of Mathematics & Statistics

Faculty of Science and Technology

Vishwakarma University, Pune

DECLARATION

We

HAMZA DALAL (202000092),

SHRAVANI UNCHE (202001320),

BHAGYALAXMI MAHARANA (202000755) &

DEEKSHA DESHMUKH (202001248)

here by declare that the work embodied in this project entitled "**FAKE JOB POSTING PREDICTION**" carried out by us under the supervision of **Prof. Mahfooz Alam, Assistant professor, Department of Mathematics & Statistics, Vishwakarma University, Pune** is an original work and does not contain any work submitted for the award of any degree in this University or any other University.

HAMZA DALAL

SHRAVANI UNCHE

BHAGYALAXMI MAHARANA

DEEKSHA DESHMUKH

M.Sc. Statistics – Big Data Analytics

Department of Mathematics & Statistics

Vishwakarma University, Pune

CERTIFICATE

This is to certify that the project of titled "**FAKE JOB POSTING PREDICTION**" submitted by **HAMZA DALAL, SHRAVANI UNCHE, BHAGYALAXMI MAHARANA and DEEKSHA DESHMUKH**, is an original work and has not been previously submitted in part or full for the award of any other degree or diploma to this or any other university. The project is submitted to **Vishwakarma University Pune**, in partial fulfilment of the requirement for the award of the degree of **Master of Science in the subject of Statistics – Big Data Analytics.**

Date: 23-May-2022

DR. NAZIA WAHID
Project Guide

PROF. MAHFOOZ ALAM
Project Guide

ACKNOWLEDGEMENT

Inspiration and Motivation have always played key roles in the success of any venture. It is our good fortune and a matter of pride and privilege to have the esteemed supervision of **Mr. MAHFOOZ ALAM, Assistant Professor, Department of Statistics, Vishwakarma University, Pune**. It is only her personal influence, expert guidance and boundless support that enabled us to complete the work, and **Dr. NAZIA WAHID, Head of the Department, Department of Statistics, Vishwakarma University, Pune** who has inculcated in us the interest and inspiration and been a constant source of motivation. We express our sincere thanks and the deep sense of gratitude towards everyone involved in any way whatsoever for the shaping of this Project.

Date: 23-May-2022

HAMZA DALAL

SHRAVANI UNCHE

BHAGYALAXMI MAHARANA

DEEKSHA DESKMUKH

Fake-Job-Posting-Prediction

1. Objective of the project:

This project aims to create a classifier that will have the capability to identify fake and real jobs. The final result will be evaluated based on two different models. Since the data provided has both numeric and text features one model will be used on the text data and the other on numeric data. The final output will be a combination of the two. The final model will take in any relevant job posting data and produce a final result determining whether the job is real or not.

2. Literature Review:

The problem of job posting identification arises due to numerous job postings which includes real jobs as well as fake jobs which are posted extensively to scam peoples or to steal identities or personal information of the applicants. Due to advancement in modern technology and social communication advertising new job posts has become very common issue in the present world. So, fake job posting prediction is a great concern to everyone. Like many other classification tasks, fake job prediction leaves a lot of challenges to face. Different data mining techniques and classification algorithm like KNN, Decision Tree, Support Vector Machine Naïve Bayes Classifier, Random Forest Classifier, etc. are used to identify real and fake jobs respectively.

The maximum ratio of fake to real job posting is of Bakersfield city of California i.e., 15:1. This means in every 16 jobs posted only 1 will tend to be genuine and rest 15 will tend to be fake. So, in this case predicting real job posting is the most difficult part as data of this real jobs would be very low as compared to the data of fake jobs.

3. Project Description:

Employment scams are on the rise. According to CNBC, the number of employment scams doubled in 2018 as compared to 2017. The current market situation has led to high unemployment. Economic stress and the impact of the coronavirus have significantly reduced job availability and the loss of jobs for many individuals. A case like this presents an appropriate opportunity for scammers. Many people are falling prey to these scammers using the desperation that is caused by an unprecedented incident. Most scammer do this to get personal information from the person they are scamming. Personal information can contain address, bank account details, social security number etc. We university student, and We have received several such scam emails. The scammers provide users with a very lucrative job opportunity and later ask for money in return. Or they require investment from the job seeker with the promise of a job. This is a dangerous problem that can be addressed through Machine Learning techniques and Natural Language Processing (NLP).

This project uses data provided from Kaggle. This data contains features that define a job posting. These job postings are categorized as either real or fake. Fake job postings are a very small fraction of this dataset. That is as excepted. We do not expect a lot of fake jobs postings. This project follows five stages. The five stages adopted for this project are –
Problem Definition (Project Overview, Project statement and Metrics)
Data Collection
Data cleaning, exploring and pre-processing
Modeling
Evaluating

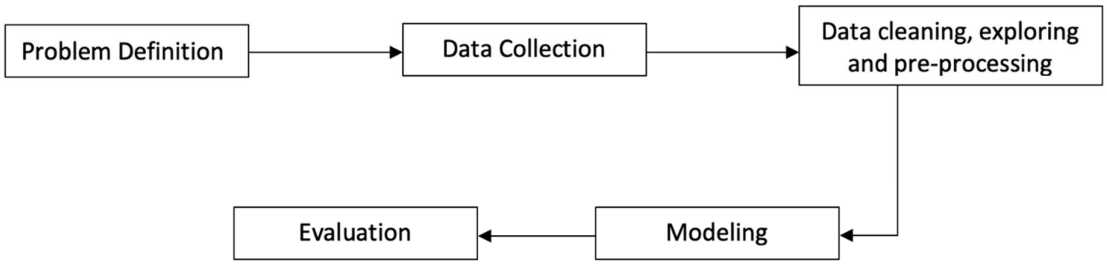


Figure 1. Stages of development

4. Tools/ Software requirements

Python
IDE: - Jupyter Notebook

Algorithms and Techniques

Based on the initial analysis, it is evident that both text and numeric data is to be used for final modeling. Before data modeling a final dataset is determined. This project will use a dataset with these features for the final analysis:

1. telecommuting
2. fraudulent
3. ratio: fake to real job ratio based on location
4. text: combination of title, location, company_profile, description, requirements, benefits, required_experience, required_education, industry and function
5. character_count: Count of words in the textual data Word count histogram

Further pre-processing is required before textual data is used for any data modeling.

The algorithms and techniques used in project are:

1. Natural Language Processing
2. Logistic Regression

5. Analysis

Data Exploration: -

The data for this project is available at Kaggle - https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction. The dataset consists of 17,880 observations and 18 features.

The data is combination of integer, binary and textual datatypes. A brief definition of the variables is given below:

#	Variable	Datatype	Description
1	job_id	int	Identification number given to each job posting
2	title	text	A name that describes the position or job
3	location	text	Information about where the job is located
4	department	text	Information about the department this job is offered by
5	salary_range	text	Expected salary range
6	company_profile	text	Information about the company

#	Variable	Datatype	Description
7	description	text	A brief description about the position offered
8	requirements	text	Pre-requisites to qualify for the job
9	benefits	text	Benefits provided by the job
10	telecommuting	boolean	Is work from home or remote work allowed
11	has_company_logo	boolean	Does the job posting have a company logo
12	has_questions	boolean	Does the job posting have any questions
13	employment_type	text	5 categories – Full-time, part-time, contract, temporary and other
14	required_experience	text	Can be – Internship, Entry Level, Associate, Mid-senior level,

#	Variable	Datatype	Description
			Director, Executive or Not Applicable
15	required_education	text	Can be – Bachelor's degree, high school degree, unspecified, associate degree, master's degree, certification, some college coursework, professional, some high school coursework, vocational
16	Industry	text	The industry the job posting is relevant to
17	Function	text	The umbrella term to determining a job's functionality
18	Fraudulent	boolean	The target variable ◊ 0: Real, 1: Fake

Since most of the datatypes are either Booleans or text a summary statistic is not needed here. The only integer is job_id which is not relevant for this analysis. The dataset is further explored to identify null values.

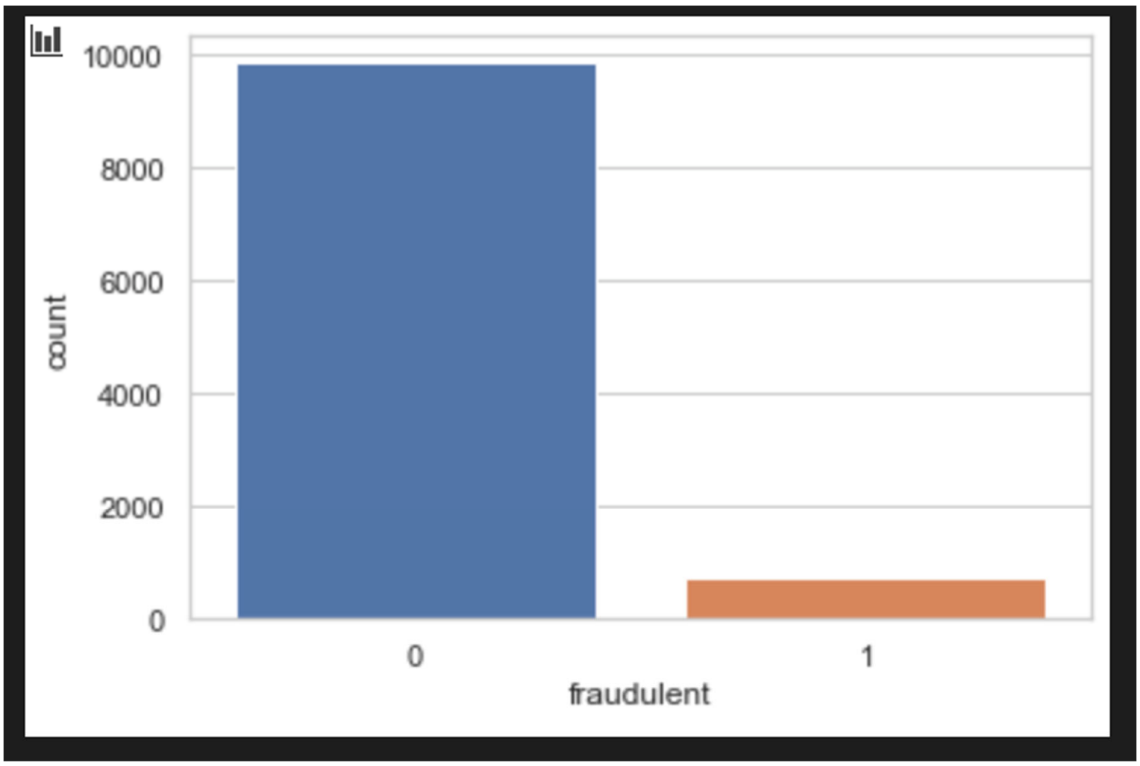
job_id	0
title	0
location	346
department	11547
salary_range	15012
company_profile	3308
description	1
requirements	2695
benefits	7210
telecommuting	0
has_company_logo	0
has_questions	0
employment_type	3471
required_experience	7050
required_education	8105
industry	4903
function	6455
fraudulent	0

Variables such as department and salary_range have a lot of missing values. These columns are dropped from further analysis.

After initial assessment of the dataset, it could be seen that since these job postings have been extracted from several countries the postings were in different languages. To simplify the process this project uses data from US based locations that account for nearly 60% of the dataset. This was done to ensure all the data is in English for easy interpretability. Also, the location is split into state and city for further analysis. The final dataset has 10593 observations and 20 features.

The dataset is highly unbalanced with 9868 (93% of the jobs) being real and only 725 or 7% of the jobs being fraudulent. A countplot of the same can show the disparity very clearly.

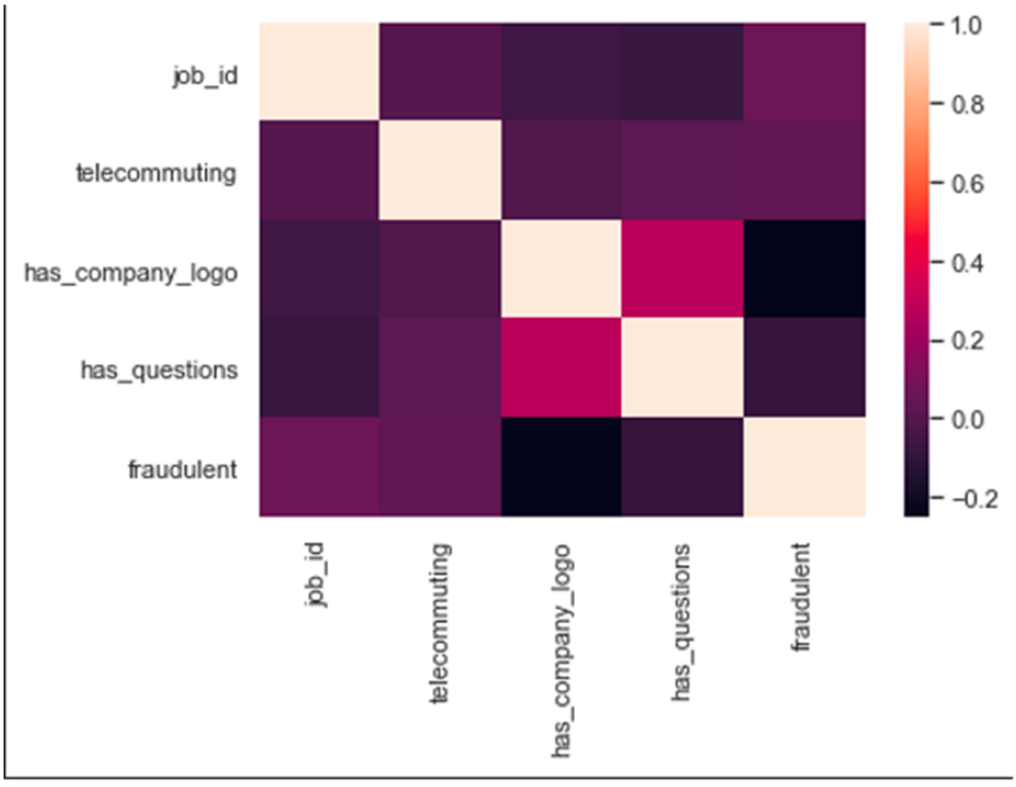
Exploratory Data Analysis :-



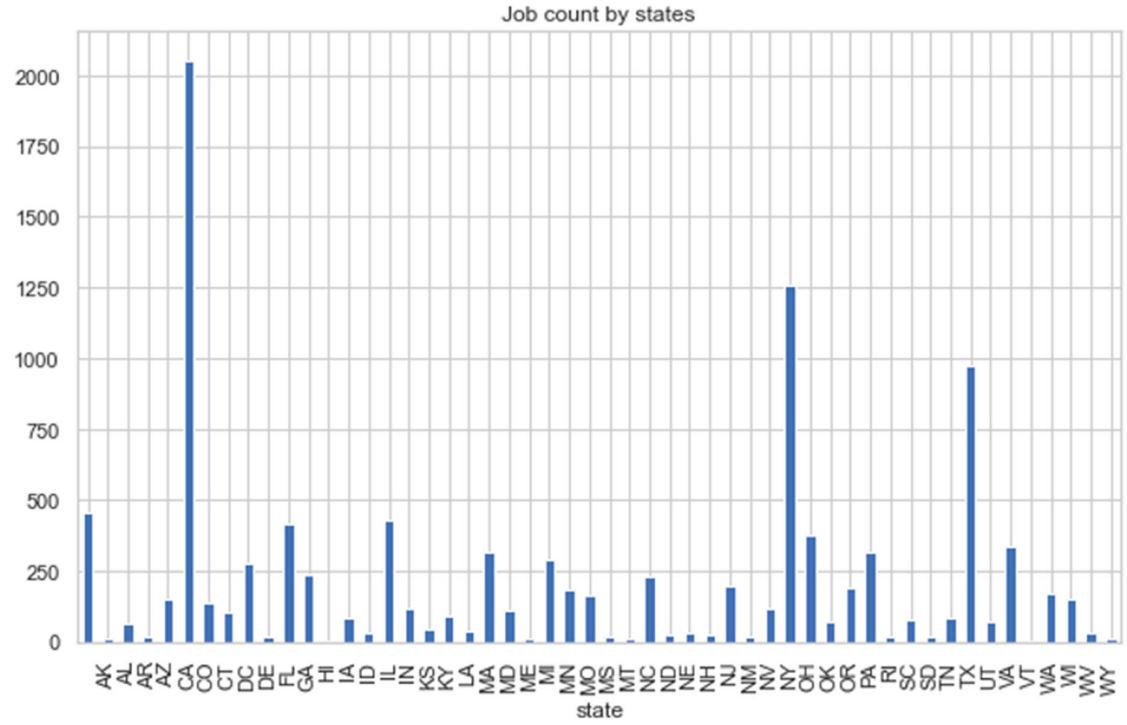
The first step to visualize the dataset in this project is to create a correlation matrix to study the relationship between the numeric data.

Figure 4. Correlation matrix

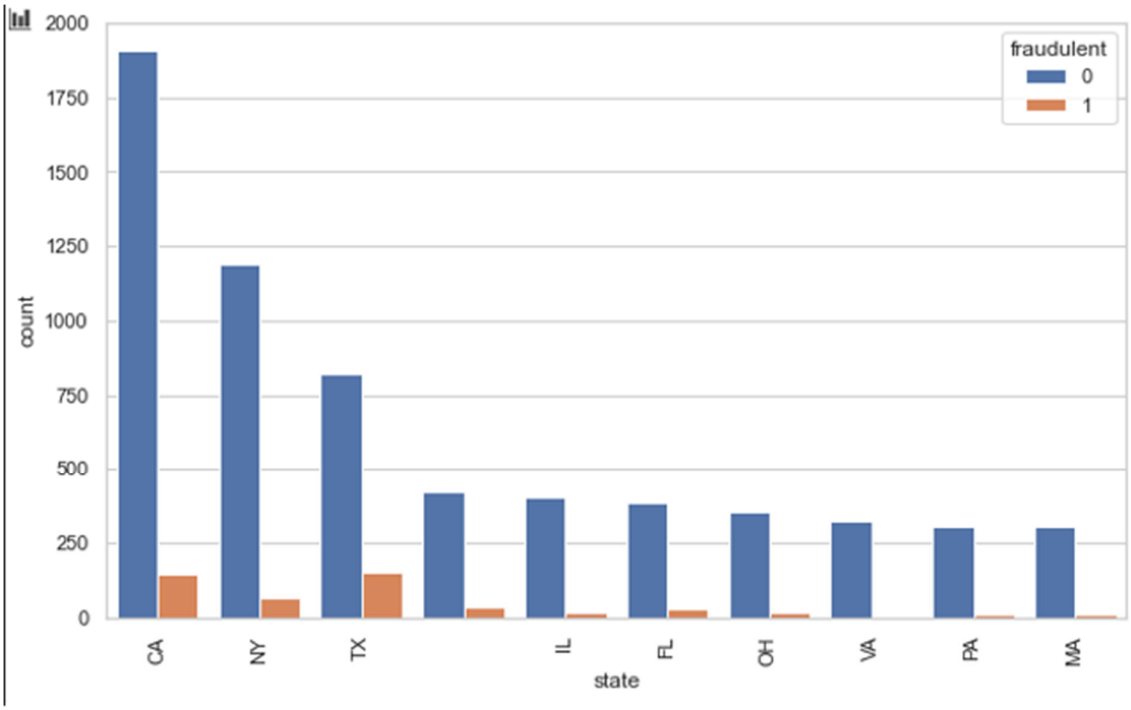
The correlation matrix does not exhibit any strong positive or negative correlations between the numeric data.



However, an interesting trend was noted with respect to the Boolean variable telecommuting. In cases when both this variable had value equal to zero there is a 92% chance that the job will be fraudulent. After the numeric features the textual features of this dataset is explored. We start this exploration from location.



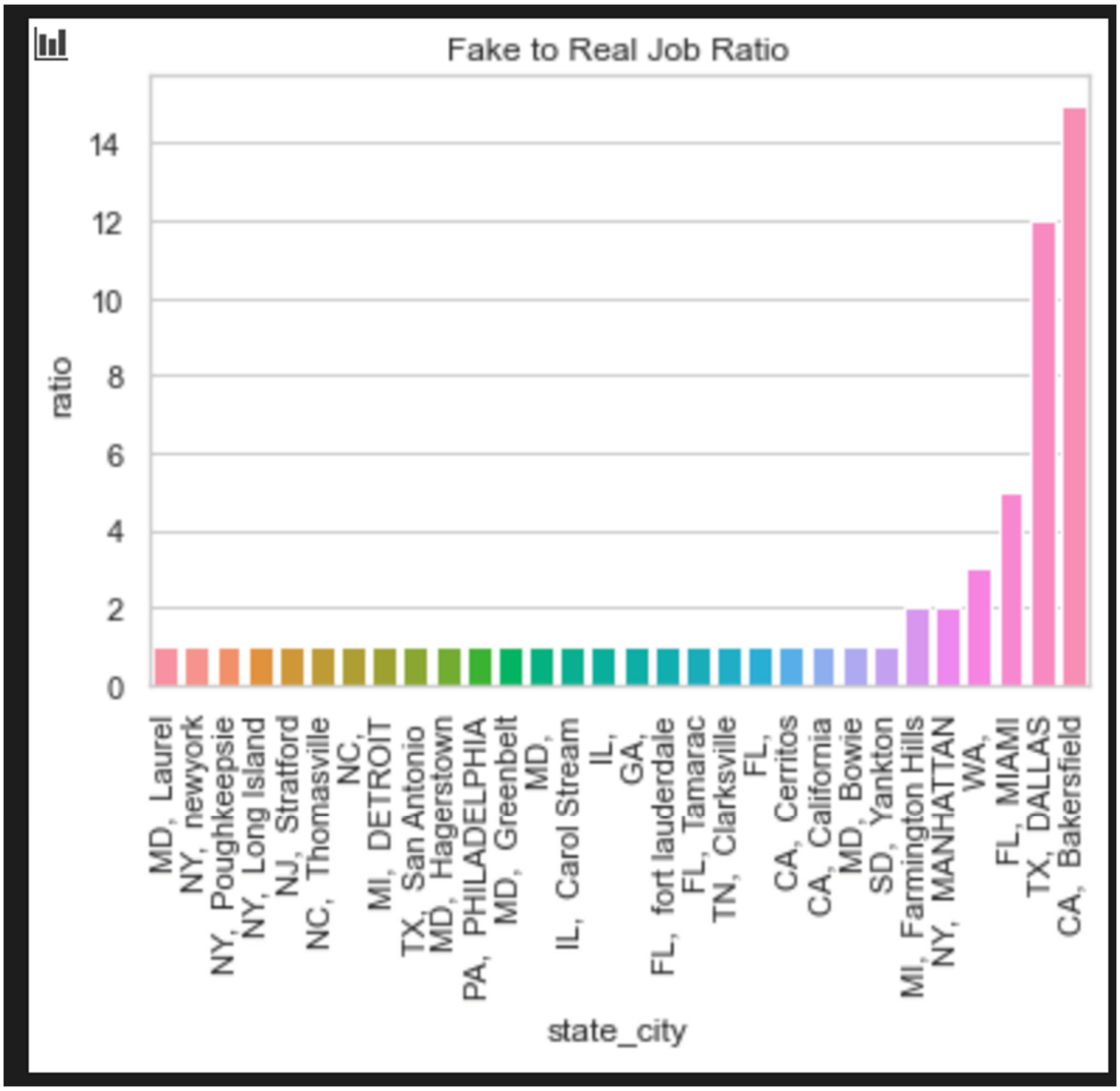
The graph above shows which states produces the greatest number of jobs. California, New York and Texas have the highest number of job postings. To explore this further another bar plot is created. This barplot shows the distribution of fake and real jobs in the top 10 states.



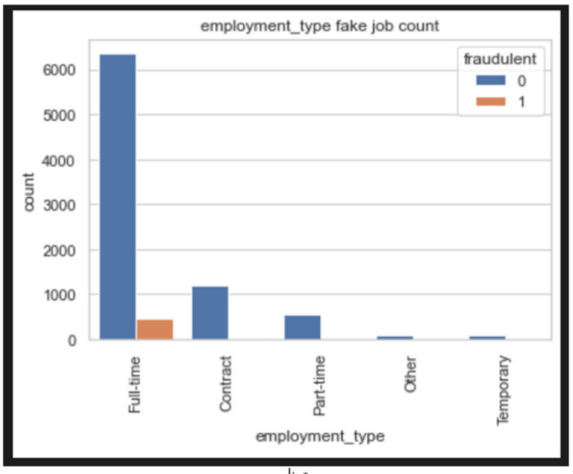
The graph above shows that Texas and California have a higher possibility of fake jobs as compared to other states. To dig one level deeper into and include states as well a ratio is created. This is a fake to real job ratio based on states and citifies. The following formula is used to compute how many fake jobs are available for every real job:

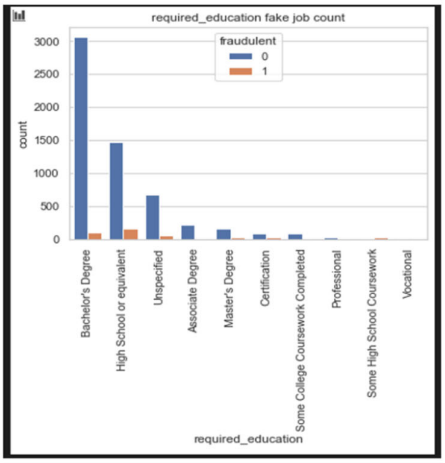
$$ratio = \frac{state\ \&\ city \mid fraudulent = 0}{state\ \&\ city \mid fraudulent = 1}$$

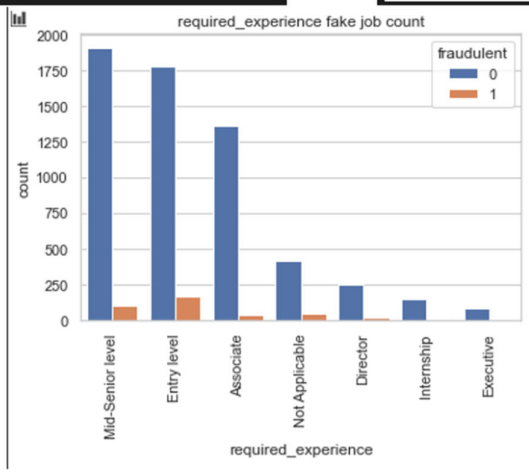
Only ratio values greater than or equal to one are plotted below.



Bakersfield in California has a fake to real job ratio of 15:1 and Dallas, Texas has a ratio of 12:1. Any job postings from these locations will certainly have a high chance of being fraudulent. Other text-based variables are explored further to visualize the presence of any important relationships.

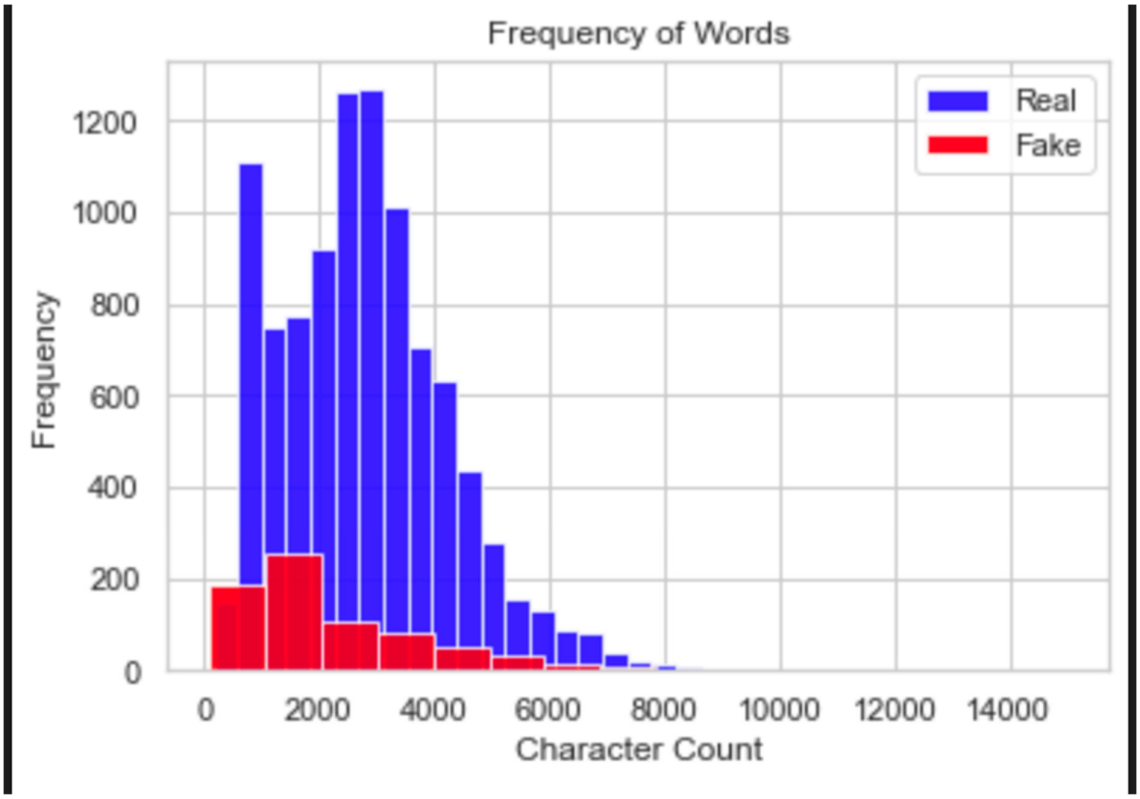






The graphs above show that most fraudulent jobs belong to the full-time category and usually for entry-level positions requiring a bachelor's degree or high school education.

To further extend the analysis on text related fields, the text-based categories are combined into one field called text. The fields that are combined are - title, location, company_profile, description, requirements, benefits, required_experience, required_education, industry and function. A histogram describing a character count is explored to visualize the difference between real and fake jobs. What can be seen is that even though the character count is fairly similar for both real and fake jobs, real jobs have a higher frequency.



6. Result :-

Observing confusion matrix we get following results:
Actual job predicted = 5104
Fake Job Predicted = 129
Fake Job Predicted as Actual = 8
Actual Job predicted as Fake = 123

7. Conclusion :-

Here we observe that Prediction of Actual Jobs is more efficient that prediction of Fake Job and that is because we have more data on Actual Jobs rather than fake jobs.

8. Future Scope of the study

- Through this series of articles, we have tried to put forward an issue that is creeping through the job market. Turmoil and chaos are the perfect proponents for scammers, and currently, cyber scam attacks are on the rise.
- We have provided a detailed analysis of how we can apply machine learning to predict the occurrences of such fake postings.
- For the ending note: Stay safe and be aware of fake job postings

9. Reference

https://ijettjournal.org/Volume-68/Issue-4/IJETT-V68I4P209S.pdf